A Review on Information Retrieval in the Historical Digital Humanities Domain

Boyang Zhang, Tampere Universities, Finland
Sanna Kumpulainen, Tampere Universities, Finland
Heikki Keskustalo, Tampere Universities, Finland

The Asian Conference on Arts & Humanities 2021
Official Conference Proceedings

Abstract
Digital humanities entail applying computational tools and methods to traditional humanities research. In this paper, we focus on history, which can be seen either as part of humanities or social sciences research. We approach the subject from the point of view of digital methods in humanities research and information retrieval. The purpose of this study is to explore the themes emerging in the recent literature concerning historians' changing work tasks in the digital era. We present a literature review based on a careful inspection of the focused sample of 47 conference/journal articles discussing digital humanities from the point of view of historical research and information retrieval. The results illustrate the requirements and needs of historians working with recent technology, the types of data discussed in the literature, and technologies and tools available to answer historians' needs. We observe and report recurring themes in the literature in order to give an overview of the subject.

Keywords: Digital Humanities, Information Retrieval, Work Tasks

Introduction

Due to various large-scale efforts focusing on digitizing historical sources, the number and type of digital historical documents available have dramatically increased (Anderson, 2004). *Information retrieval* (IR) from digital historical sources can provide historians with valuable access to a significant amount of historical information, and it is a vast and diffuse field of study. Our motivation in this paper came from our desire to understand what kinds of issues the recent literature deals with regarding the digital humanities, historians' work, and information retrieval. This paper's point of departure is the following overall research question: *what kinds of themes emerge* in the recent scientific literature discussing *historians' information retrieval in the digital era*? We elaborate this general question by articulating the following sub-questions (RQ1-RQ3):

- RQ1: What kinds of *requirements and needs* of historians working with digital sources are reported in the literature?
- RQ2: What *types of data* are discussed in the literature?
- RQ3: What kinds of modern *technologies and tools* are utilized in the literature?

Method

To conduct a systematic literature review, we utilized Fink's process model entailing (Fink, 2005) seven stages. Our *steps* included: (1) selecting the research question; (2) selecting the database(s) to search papers for the review; (3) defining the search terms; (4) setting the primary screening criteria to exclude irrelevant papers and include relevant papers; (5) applying the criteria; (6) reviewing the included papers; (7) presenting the synthesis of the results.

A growing number of IR researchers are discussing evaluating methods and experimenting with technologies/tools. Kelly et al. (2013) review the interactive information retrieval systems (1967-2006) which begin to innovate in methods and algorithms. Continuously, our overall goal is to get a grasp of the themes emerging in recent literature regarding IR in the historical digital humanities domain. We desire to find out about the different viewpoints related to historians' work in digital environments. To facilitate our exposure to various viewpoints, we decided to use a two-fold strategy. First, we harvested a large set of papers limiting ourselves to high-quality sources and using a systematic query approach combining the main concepts focusing on our subject of interest. Admitting that many relevant documents might not match our search expression (due to their specific topic or level of abstraction), we augmented the initial set of articles retrieved by additional papers based on intellectual searching effort. Subsequently, we gathered a set of papers that collectively shed light on information retrieval in historians' work tasks in the digital era.

Regarding the first step (see above), our overall research question was to find out what kinds of themes emerge in recent scientific literature when we focus on IR in the historical digital humanities domain. In the second step, three databases were selected for searching, in our case EBSCO, ProQuest and Scopus. In the third step, keywords were selected to describe the concepts of "history" or "historical", "digital humanity" or "digital humanities", and "information retrieval" related research. We decided to utilize the query "histor* OR (digital humanit*)" AND "information retrieval" after several tryouts in EBSCO and ProQuest databases. Based on the result set analysis, this query facilitated exposure to a variety of viewpoints taken in the studies, while simultaneously keeping the result set reasonably small.

Similarly, in Scopus database, we used the query ALL((histor* AND (digital humanit*)) AND (information retrieval)) AND (LIMIT-TO (LANGUAGE,"English")).

Regarding steps four and five, our selection criteria included the sole publication language *English*; full texts of peer reviewed conference or journal articles were accepted. The third-party citation software Zotero was applied to download the references for setting up a data extraction table. The period of the publications was restricted from 1997 to 2017, without other constraints. Regarding step six, the first author inspected abstracts regarding 98 papers retrieved. During this phase five duplicate articles and 54 non-relevant papers were detected (See Table 1). Duplicates and non-relevant papers were removed from the original set of 98 documents. The remaining set of 44 relevant papers was supplemented by 3 relevant papers acquired based on snowballing technique used to retrieve strong related papers (Wohlin, 2014). The remaining 47 papers were read by the first author in their entirety, and a data extraction table was created based on reading. Our conceptual analysis in the remainder of this paper is based on this set of 47 documents.

Database	**Number of articles**
Size of the document set retrieved from EBSCOhost (searched 02.05.2018)	23
Size of the document set retrieved from ProQuest (searched 02.05.2018)	39
Size of the document set retrieved from Scopus (searched 02.05.2018)	41
Duplicate articles observed in the search results of EBSCOhost, ProQuest and Scopus (inspected 03.05.2018)	5
Number of distinct articles after checking the abstracts and the titles	98
Number of documents *excluded* based on the inspection of titles, abstracts and full texts.	54
Data sample included for the review after reading the whole articles	44
Additional articles retrieved from the other sources (Google Scholar)	3
Number of articles in the final set selected for the review	47

Table 1: Number of Documents in Search Results and During the Screening and Reviewing Processes

Next, a data extraction table was created to facilitate the analysis of the articles and construct a summary of each individual paper. The design of this table enabled us to categorize relevant factors observed in the articles. The articles were first ranked by timestamp from the latest to the oldest. Then, each article was intellectually analyzed by the first author, and placed into one central or core category (*one-classification* scheme of qualitative analysis) (Suter, 2012). As the seventh (and the final step), the articles' analyses and results will be presented in the next section. They are based on the data extraction table and illustrate the themes emerging in the recent scientific literature discussing historians' IR in the digital era. In this study, the creation of the classification schemes used in the analyses was based on the first author's

interpretations while reflecting the major conceptual aspects present in the papers from the research questions' point of view. Admittedly, it would be possible to create also slightly different schemes based on the same data, e.g., by selecting the level of the conceptual granularity used in the analyses differently. In the next chapter, the classification of the articles (see Tables 2 and 3) is based on the data extraction table explained above. In both tables each article is placed into one central category of the one-classification scheme based on the first author's intellectual selection.

Result

In relation to historians' IR in digital collections, various user requirements and needs, types of data, tools and technologies were observed. We were interested in seeing how these themes are discussed (in a conceptually more detailed level) in the set selected based on the systematic review. Our article selection method clearly facilitated exposure to a large variety of topics yet having a shared conceptual core (digital humanities, history, and information retrieval). Using modern tools and technologies can be beneficial for the historians' tasks, but the user task must be understood first. Therefore, in the following subsections, we will first unveil what kinds of user requirements and needs were discussed in the literature. Each paper was categorized into *one* class based on the user requirements discussed in the paper (Table 2). In case several dimensions of the user requirements were observed in the paper, the author used his own best judgement to select the most pertinent one. Secondly, this analysis is followed by describing the types of data, and tools and technologies discussed in the reviewed literature. Each paper was also categorized into *one* class, but this time based on the combination of the data type and the technologies/tools used. These final two one-class classification schemes were formed to observe relevant themes in the articles during their intellectual inspection. Table 3 explicates the technologies and tools, in connection to the data types discussed in the papers.

1. Users' requirements and needs

In 11 of the 47 papers, the theme of user-centered solutions to *accessibility* of historical information was discussed, especially from the point of view of *navigating* various types of data. Furthermore, *simplicity* was often brought into focus, related to offering clear and simple solutions including small size vocabularies and detailed information which can help historians to get exactly what they need (11 out of 47 articles). 8 out of 47 articles discussed the *interpretation* of the documents, the requirement of transparency, and the clarity of complex historical documents. *Standardization* is needed to de-construct and de-compose various evaluating criteria (7 out of 47 articles) to deal with issues such as historical astronomical resources and event timelines. Additionally, various approaches have been used to ensure *accuracy* related to historical information retrieval (7 out of 47). There were also a few studies discussing the *efficiency* (3 out of 47).

Users' requirements and needs	Number of articles	Source
Accessibility	11	(Anderson, 2004; Clifford et al., 2016; Coleman, 2006; Gregory & Schwartz, 2009; Huang & Soergel, 2006; Kemp, 2009; Kunz, 2007; Mackenzie et al., 2009; Schwartz, 2015; Smiraglia, 2003; Wilson, 2009)

Simplicity	11	(Hinrichs et al., 2015; Isoda et al., 2009; Jänicke et al., 2015; Lin et al., 2008; Matei, 2009; Porter, 2006; Saiti & Prokopiadou, 2008; Ullyot, 2014; Uzwyshyn, 2007; Yang & W. K. Luk, 2003; Zeng et al., 2014)
Interpretation	8	(Beatrice et al., 2017; Erjavec, 2015; Joint, 2009; Jones et al., 2001; Mizzaro, 1997; Thiel et al., 2004; Webb et al., 2017; Wettlaufer et al., 2015)
Standardization	7	(Cole & Leide, 2003; Garfield, 2009; Lifante & Madrid, 2015, 2015; Othman & Salahuddin, 2015; Rodríguez et al., 2017; Shabajee et al., 2006)
Accuracy	7	(Heuser & Le-Khac, 2011; Jarlbrink & Snickars, 2017; Järvelin et al., 2016; Kettunen et al., 2016; Kramer et al., 2011; McEathron, 2002; Read et al., 2016)
Efficiency	3	(Petrelli & Clough, 2012; Saracevic, 2008; Wiesman et al., 2004)

Table 2: Dimensions of Users' Requirements and Needs Discussed in the Articles Reviewed

2. Types of data and technologies/tools

This subsection analyzes the articles from the point of view of the data types and technologies/tools discussed. A considerable number of articles (36 out of 47) studied solely *digital* documents, while 11 articles addressed digital and non-digitalized documents.

The articles' common theme was to address classification, categorization, *and ranking* (14 out of 47). Combinations of different document types as *multimedia* were often discussed (10 out of 47 articles) including formats such as text and video; text and audio; text and image; text and GIS; or a combination of various digital resources. The usage of more than one document type has been considered to be a worthy character trait in historical research. *Citation index and annotation*, which include commenting, explaining and interpreting information, was a topic of seven (7) articles. Similarly, the analysis of *Geographic Information System* (GIS) was the topic of seven articles. Visualization techniques attracted relatively less attention with 5 articles. The rest of the papers handled audio- and multi-language-related topics.

Functions	Types of data	Technologies and tools
Classification, categorization, ranking 14 out of 47	Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts Digital texts	Standardization of historical astronomical resources (Rodríguez et al., 2017) Segmentation tools at digitization Zissor (Jarlbrink & Snickars, 2017) OCRed errors analysis tools (Kettunen et al., 2016) ToTrTaLe pipeline tool with tokenisation, transcription (Erjavec, 2015) Metadata Core Categories (Steiner & Koch, 2015) Standardize information retrieval parameters (Lifante & Madrid, 2015) Topic model through linguistic categories (Ullyot, 2014) CALIS and CDSSP library IR tools (Zeng et al., 2014)

		IR systems retrievals with user relevance (Saracevic, 2008) An algorithm for suffix stripping (Porter, 2006) Postmodern way of catalog (Smiraglia, 2003) Cartographic work for new mechanisms IR (McEathron, 2002) Machine translation CLIA systems (Jones et al., 2001) Chronological order ranking of documentation (Mizzaro, 1997)
Combinations of different types (text and video, multimedia, and so on) 10 out of 47	Digital texts and locations Digital texts and others Digital images Digital multimedia Digital texts and libraries Digital multimedia Digital texts and images Digital texts and networks Digital multimedia Digital texts and records	Text mining with commodities and geographical location (Clifford et al., 2016) WissKI connect both texts and other forms of objects (Wettlaufer et al., 2015) Image filtering techniques (Kramer et al., 2011) Robust inference platform with multiple sources (Lin et al., 2008) Historical & statistical database of libraries' online catalogs (Saiti & Prokopiadou, 2008) Multimedia information visualization system (Uzwyshyn, 2007) Tools for historians to make use of online retrieval (Anderson, 2004) Graphical representation of meta-information on documents (Wiesman et al., 2004) Historical multimedia collection (Shabajee et al., 2006) Topic with speech evidence (Huang & Soergel, 2006)
Citation index, annotation (Interpretation) 7 out of 47	Digital texts Digital newspapers Digital texts Digital texts Digital library Digital texts Digital films	Curation automatic process of data with manual checking (Beatrice et al., 2017) Query method to index words (Järvelin et al., 2016) Translating and visualizing data to interpretable forms (Heuser & Le-Khac, 2011) Citation analysis tools towards bibliometric data (Garfield, 2009) Google digitized book library (Joint, 2009) Modern bibliography, cataloging, classification, & indexing (Coleman, 2006) Bask-based interfaces for index and annotating_COLLATE (Thiel et al., 2004)
GIS (Geographic Information System) 8 out of 47	GIS data GIS data GIS data GIS data GIS data GIS data GIS data GIS data	Historical GIS and spatial history (Schwartz, 2015) Automatic GIS generation footprint polygons (Isoda et al., 2009) GIS with SDI (Spatial data infrastructures) (Wilson, 2009) Database, GIS contribute to historical study (Mackenzie et al., 2009)

		GIS tools to explore and visualize historical events (Kemp, 2009) Historical GIS tools (Gregory & Schwartz, 2009) Visualization tools of geographic locations (Matei, 2009) HGIS historical GIS info (Kunz, 2007)
Visualization 4 out of 47	Digital documents Digital texts Digital texts Digital texts	Historical mapping and visualization tools (Read et al., 2016) Text mining and information visualization tool (Hinrichs et al., 2015) Text Re-use Alignment Visualization TRAViz (Jänicke et al., 2015) Visualization scheme for key words (Cole & Leide, 2003)
Multi-language, 3 out of 47	Digital texts Digital library (image) Digital texts	Index Islamicus on Islamic History and Civilization (Othman & Salahuddin, 2015) CLIR cross-language IR tools (Petrelli & Clough, 2012) Multi-language interpretation (English/Chinese) (Yang & W. K. Luk, 2003)
Audio 1 out of 47	Digital Audio	Semantic analysis via audio analysis techniques (Webb et al., 2017)

Table 3: Types of Data and Technologies/Tools

Primary sources refer to the original documents which may contain large amount of number of fragmented information (Anderson, 2004; Clifford et al., 2016; Saiti & Prokopiadou, 2008). Modern technologies and tools provide certain access orders and logic sequences for historians to retrieve the information needed which can be regarded as secondary source (Coleman, 2006). In table 3, a list of tools discussed in the context of historical information retrieval is presented as more specific functionalities. They offer various systematical methods to retrieve often relatively disordered and scattered documents. The digital documents may include many types of digitalized media contents, including texts, videos, audios, images, newspapers, libraries, multimedia, GIS and so on. Table 3 classifies the technologies/tools which were used in the literature reviewed by functionalities. Our purpose is to illustrate the overall point of view in order to understand what types of data were discussed in the historical information retrieval literature and what kinds of modern technologies and tools were utilized in these papers. Regarding the research questions (see Section 1), Table 3 combines answers to RQ2 and RQ3 .

Classification, categorization, ranking. Articles were selected into this classification category based on the functionalities of the technologies and tools discussed in the papers. Rodríguez et al. (2017) propose the standards of elements for cataloguing descriptive historical astronomical resources. Jarlbrink & Snickars (2017) describe the problems emerging during transforming original historical prints through segmentation and optical character recognition into digital form. Kettunen et al. (2016) investigate the effectiveness of named entity (NE) recognition from historical text. Erjavec, T. (2015) studies the collection of Slovene historical texts with pipeline tool for teaching purposes; the pipeline tools provide tokenization, transcription, tagging and lemmatization for documents with added in-line annotation. Steiner & Koch (2015) provide metadata core categories and acquisition rules by analyzing text materials and images, museum objects and artifacts. Lifante & Madrid. (2015) digitalize and standardize

information retrieval parameters on a considerable amount of historical information stored in non-computerized formats. Ullyot (2014) provides a topic model of machine-readable transcriptions to simplify historical linguistic categories. Zeng et al. (2014) illustrate scarce sources as one-step access as document supply service platform to retrieve literature resources in China Academic Library. Smiraglia (2003) illustrates postmodern catalogs to explore the informative capability of works. McEathron (2002) studies new mechanisms for using historical cartographic works as entities for information retrieval. Jones et al. (2001) shed light on machine translation and cross-language information tools to overcome language barriers between the history of English and Japanese.

Combinations of different types. Articles were selected into this category to include technologies and tools dealing with multiple media types. Clifford et al. (2016) identify the relationships between texts and commodities, geographical locations and dates to map the changing geography. Wettlaufer et al. (2015) explore the relationships between texts and museum objects with semantic web technologies to support user experiences. Kramer et al. (2011) process the analog land-use maps to digital European historical land-use database to enhance the image filter techniques. Lin et al. (2008) investigate the robust inference platform for real-life knowledge discovery and integration over different distributed sources. Saiti & Prokopiadou (2008) examine internet as post-graduate students' primary information source and how to perform fast information retrieval from the combination of historical and statistical database. Uzwyshyn (2007) studies the multimedia visualization and interactive systems to connect sider spectrum of media elements. Anderson (2004) provides modern tools on classification and categorization for historians to make use of online retrieval. Wiesman et al. (2004) present the concept of metabrowing to present networkswhich are related to the digital document contents of different types. Shabajee et al. (2006) develop a prototype for digital resource discovery portal from historical multimedia collection with semantic web technologies. Huang & Soergel (2006) investigate the relevance between speech and topic note, and make connections between evidence and a topic.

Citation index, annotation. Articles were selected into this category based on the technologies and tools utilizing citation index and annotation methods. Beatrice et al. (2017) evaluate large-scale digital and literary documents to assist the automatic process of extensive data with manual checking. Järvelin et al. (2016) presents a traditional test collection-based evaluation on the effectiveness of information retrieval, using fuzzy string matching methods in generating query expansion terms for retrieving historical documents written in a highly inflectional compound language (Finnish). Heuser & Le-Khac (2011) bring out ways of translating and visualizing data into readily interpretable forms. Garfield (1998) discusses the retrieval of related bibliometric data from citation analysis. Joint (2009) studies the settlement of Google's digitized books, and discusses various related philosophical and moral issues. Coleman (2006) considers user-centered way of thinking regarding designing digital information organizations and services. Thiel et al. (2004) investigate the designing of content and content-based knowledge working environment for distributed user groups to work with digital document sources.

GIS. Articles were selected into this category based on the technologies and tools dealing with geographic information system (GIS) information. Schwartz (2015) combines text mining and GIS to bring out spatial relationships on any kind of documents. Isoda et al (2010) develop an automatic tool to generate realistic virtual reality (VR) models based on GIS data to have easy access to virtual space. Wilson (2009) uses spatial data infrastructures to develop GIS standards that can be globally accepted. Mackenzie et al. (2009) discuss the usage of GIS to provide a

mapping function via Web Map services. Kemp (2009) discuss GIS in the context of exploring and visualizing historical events. Gregory & Schwartz (2009) propose a tool on historical GIS to better understand the past's geographies. Matei, S. (2009) discusses the multi-dimensional tools to visualize geographic realities. Finally, Kunz (2007) studies GIS data concerning Germany's states and territories during the nineteenth century.

Visualization. Articles were chosen into this category based on including technologies and tools dealing with visualization methods. Read et al. (2016) analyze the importance of technologies in modern historical researches to advance humanitarian information systems. Hinrichs et al. (2015) combines text mining tools and visualization method in large-scale environmental history to study commodity trade. Jänicke et al. (2015) provide text re-use alignment visualization tool to assist users' engagement on historical and modern texts. Coleman (2006) considers user-centered way of thinking in designing digital information organizations and services.

Multi-language. Articles were categorized into this group based on discussing technologies and tools dealing with multiple languages. Othman & Salahuddin (2015) measure the relevance status of index salamicus on Islamic history and Vicilization to rank documents and provide better way of indexing. Petrelli & Paul Clough. (2012) develop cross-language information retrieval prototype on Italian to English image retrieval system with studies on user's search behaviors. Yang & Luk (2003) show the automatic thesaurus tool to retrieve multi-language background documents with interpretation.

Audio. This final group includes the articles focusing on audio technologies and tools. Webb et al. (2017) interpret audio files by generic audio analysis methods to extract semantic information from digital audio files.

Discussion

Requirements and needs. To answer the question regarding the requirements and needs of historians, the selected literature was categorized into accessibility; simplicity; interpretation; standardization; accuracy and efficiency (Table 3). From the users' perspective, accessibility is required by historians to perform information retrieval. The IR tools design could engage users in planning and adopt the search items satisfying users' requirements and needs (Anderson, 2004; Clifford et al., 2016; Coleman, 2006; Huang & Soergel, 2006; Smiraglia, 2003). Moreover, the significance of simplicity of information retrieval equally requires attention. Can historians retrieve the needed information simply? Research in user-centered information retrieval has begun to depict specific requirements and needs of historians, related to issues such as events, dates, and gender (Hinrichs et al., 2015; Isoda et al., 2009; Saiti & Prokopiadou, 2008). Apparently, the retrieved items need to be interpreted according to the requirements and needs of the users. Moreover, the context in which the items are interpreted is important (Beatrice et al., 2017; Erjavec, 2015; Joint, 2009; Jones et al., 2001). The sources of retrieval and documents should be standardized by utilizing common parameters. Standard information retrieval tools can be provided for historians, by utilizing applications such as PHP, HTML, GIS, CSS, WWW. Yet, also the accuracy and efficiency should be enhanced in the design of information retrieval tools for historians to satisfy their requirements and needs.

Types of data. Base on this literature review, it is obvious that information retrieval has exerted a profound impact on modern history- related research. Admittedly, the utilization of digital format has become a tendency (Heuser & Le-Khac, 2011; Jarlbrink & Snickars, 2017; Matei,

2009; Schwartz, 2015; Shabajee et al., 2006). Digital documents have gradually become a preferred form by the historians (Anderson, 2004; Case, 1991; Kettunen et al., 2016). The retrieval of digital historical documents has fundamentally changed the way historians work. On the other hand, digitization has greatly improved the quality and speed of historical research (Kunz, 2007; Read et al., 2016).

Technologies and tools. Because the historical *primary* sources may consist of un-ordered raw data, to help working with this type of data, workflows (Joint, 2009); standards (Anderson, 2004; Coleman, 2006; Rodríguez et al., 2017; Saracevic, 2008); and ranking methods (McEathron, 2002; Mizzaro, 1997; Smiraglia, 2003) have been developed and discussed in the literature. At the same time, the significant processing challenges related to noisy data have been in the focus of many studies (Beatrice et al., 2017; Järvelin et al., 2016; Kettunen et al., 2016). Subsequently, methods and tools to deal with these challenges have been developed and critically discussed, for example, related to the segmentation and tokenization processes (Jarlbrink & Snickars, 2017).

The relationships between various media types also has been demonstrated in the literature (Lifante & Madrid, 2015). Examples of various points of views include combinations between texts, commodities, geographical location and dates (Clifford et al., 2016); texts, historical photographs and lantern slides, museum objects (Steiner & Koch, 2015); and texts and objects in general (video, audio, image, datasets, etc) (Lin et al., 2008; Saiti & Prokopiadou, 2008; Shabajee et al., 2006; Wettlaufer et al., 2015; Wiesman et al., 2004; Zeng et al., 2014).

Many studies discussed topics related to citation indexes and annotations. The readers can search and retrieve relevant topics by examining the cited references which are regarded by means of citation analysis (Garfield, 2009).Annotations are able to illustrate a content and context-based workflows and the knowledge of users (Thiel et al., 2004); locate the relevance between material and topics (Huang & Soergel, 2006; Othman & Salahuddin, 2015); translate and interpret to interpretable forms (Heuser & Le-Khac, 2011) and machine-readable forms (Ullyot, 2014) which can be applied to build connection with other RDF data models. For example, Erjavec (Erjavec, 2015) suggested a pipeline tool to perform tokenization and transcription for adding in-line linguistic annotations.

Finally, studies related to GIS (Geographic Information System) related topics were continuously carried out in relation to the development of digital documents, especially the development of historical GIS and spatial history (Gregory & Schwartz, 2009; Isoda et al., 2009; Kemp, 2009; Matei, 2009; Schwartz, 2015; Wilson, 2009); and, for example, mapping and locating historical land-use classes (Kramer et al., 2011; Kunz, 2007). Similarly to geo-location, visualization schemes have attracted a growing interest in recent years. For example, we may mention the topics of keywords visualization on specific topics (Cole & Leide, 2003) or variation between editions of both historical and modern texts (Jänicke et al., 2015); visual technology and crisis mapping using social media and SMS data (Read et al., 2016); environmental history exploration by text mining and information visualization on commodity trade (Hinrichs et al., 2015). Nevertheless, multi-language related information retrieval also brings the attention from researchers (Jones et al., 2001; Petrelli & Clough, 2012; Yang & W. K. Luk, 2003). Mistranslations can directly influence the effectiveness of the retrieval results.

Conclusion

In this paper we presented a systematic literature review on the topic of IR in the historical digital domain. Our goal was to explore the themes emerging in the recent scientific literature discussing this topic. We conducted a systematic literature review based on the Fink's seven-step model. After querying three high quality databases, and performing the initial screening, 47 articles were selected for in-depth scrutiny. These papers discussed the topic from various viewpoints, thereby facilitating our exposure to a broad range of themes.

In conclusion, this research constructs an analytical framework in order to scrutinize information retrieval in the historical domain and understand digital historians' requirements and needs. The findings of this study indicate the types of methodological tools and procedures to support the historians performing work tasks. By identifying the historians' requirements and needs, it is vitally significant to locate appropriate information by using various information retrieval methods. The analysis was conducted to explore the requirements and needs of historians from diverse perspective in order to identify relevant factors discussed. Our analytical framework contained the following major elements to guide our attention while inspecting the papers:

1. Addressing the *requirements and needs* of historians.
2. Understanding on *types of data*;
3. Modern *technologies and tools* used.

Our system of analysis helps systematically address how modern information retrieval tools have been used to access historical information to satisfying historians' requirements and needs.

Advocating for meeting the *requirements and needs* of historians is vitally significant in designing information retrieval tools. The individual work tasks may vary significantly in history related tasks. Better understanding of various types of data objects (texts, images, audio, and video) allows historians to access and interpret the historical information with simplicity. In the process of historical information retrieval, utilization of standards may allow, e.g., the identification of common types of searches applied, and unified criteria. By taking advantage of technology, historians may access the valuable data more accurately and efficiently.

The second element was to understand *the types of data*. This paper focused on digital documents, which are machine-readable and computer-readable. Historians are able to read, store, transfer and retrieve digital documents. Corresponding to different types of data, different ways to analyze exist In case of textual documents, text mining techniques and Natural Language Processing (NLP) tools can be used to analyze the features o of the texts and to identify patterns of the text. Correspondingly, image recognition tools can be used to analyze images which contains historical photographs, such as images of lantern slides.

Moreover, video clips can be considered as a group of images composed together, with sound added. Specific methods exist to process information in audio documents. Last, based on analyzing Geographic Information System (GIS) data, it is possible to build geographic realities and 3D models in the historical domain and, e.g., visualize historical changes. Also, combination methods can be used to de-compose multiple types of data, for example, to manage multimedia existing in digital libraries (Uzwyshyn, 2007).

This study aimed to provide a general view regarding the aspects of interest discussed recently in the domain of information retrieval in the historical digital humanities domain. In the future studies, the focus could be pointed towards more specific issues and further investigated, for example, how information retrieval tools can benefit more specific needs of the users accessing historical documents, and how to enhance the actual *task* performance based on understanding the practical requirements and actual needs of specific historians.

References

Anderson, I. G. (2004). Are You Being Served? Historians and the Search for Primary Sources. Archivaria, 58(0). https://archivaria.ca/index.php/archivaria/article/view/12479

Beatrice, A., Grover, C., Oberlander, J., Thomson, T., Anderson, M., Loxley, J., Hinrichs, U., & Zhou, K. (2017). Palimpsest: Improving assisted curation of loco-specific literature. *Digital Scholarship in the Humanities,* 32(suppl_1), i4–i16. https://doi.org/10.1093/llc/fqw050

Case, D. O. (1991). Conceptual organization and retrieval of text by historians: The role of memory and metaphor. *Journal of the American Society for Information Science,* 42(9), 657–668. https://doi.org/10.1002/(SICI)1097-4571(199110)42:9<657::AID-ASI4>3.0.CO;2-7

Clifford, J., Alex, B., Coates, C. M., Klein, E., & Watson, A. (2016). Geoparsing history: Locating commodities in ten million pages of nineteenth-century sources. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 49(3), 115–131. https://doi.org/10.1080/01615440.2015.1116419

Cole, C., & Leide, J. E. (2003). Using the user's mental model to guide the integration of information space into information need. *Journal of the American Society for Information Science and Technology,* 54(1), 39–46. https://doi.org/10.1002/asi.10172

Coleman, A. S. (2006). William Stetson Merrill and bricolage for information studies. *Journal of Documentation,* 62(4), 462–481. https://doi.org/10.1108/00220410610673855

Erjavec, T. (2015). The IMP historical Slovene language resources. *Language Resources and Evaluation,* 49(3), 753–775. https://doi.org/10.1007/s10579-015-9294-7

Fink, A. (2005). *Conducting Research Literature Reviews: From the Internet to Paper.* SAGE.

Garfield, E. (2009). From Citation Indexes to Informetrics: Is the Tail Now Wagging the Dog ? *Libri,* 48(2), 67–80. https://doi.org/10.1515/libr.1998.48.2.67

Gregory, I., & Schwartz, R. M. (2009). National Historical Geographical Information System as a tool for historical research: Population and railways in Wales, 1841–1911. *International Journal of Humanities and Arts Computing,* 3(1–2), 143–161. https://doi.org/10.3366/ijhac.2009.0013

Heuser, R., & Le-Khac, L. (2011). Learning to Read Data: Bringing out the Humanistic in the *Digital Humanities*. Victorian Studies, 54(1), 79–86.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., & Coates, C. M. (2015). Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities,* 30(suppl_1), i50–i75. https://doi.org/10.1093/llc/fqv046

Huang, X., & Soergel, D. (2006). *An evidence perspective on topical relevance types and its implications for exploratory and task-based retrieval* [Text]. http://www.informationr.net/ir/12-1/paper281.html

Isoda, Y., Tsukamoto, A., Kosaka, Y., Okumura, T., Sawai, M., Yano, K., Nakata, S., & Tanaka, S. (2009). Reconstruction of Kyoto of the Edo era based on arts and historical documents: 3D urban model based on historical GIS data. *International Journal of Humanities and Arts Computing,* 3(1–2), 21–38. https://doi.org/10.3366/ijhac.2009.0007

Jänicke, S., Geßner, A., Franzini, G., Terras, M., Mahony, S., & Scheuermann, G. (2015). TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities,* 30(suppl_1), i83–i99. https://doi.org/10.1093/llc/fqv049

Jarlbrink, J., & Snickars, P. (2017). Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. *Journal of Documentation,* 73(6), 1228–1243. https://doi.org/10.1108/JD-09-2016-0106

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., & Kettunen, K. (2016). Information Retrieval from Historical Newspaper Collections in Highly Inflectional Languages: A Query Expansion Approach. J. Assoc. *Inf. Sci. Technol.,* 67(12), 2928–2946. https://doi.org/10.1002/asi.23379

Joint, N. (2009). The Google Book settlement and academic libraries. *Library Review,* 58(5), 333–340. https://doi.org/10.1108/00242530910961756

Jones, G., Collier, N., Sakai, T., Sumita, K., & Hirakawa, H. (2001). A Framework for Cross-Language Information Access: Application to English and Japanese. *Computers and the Humanities,* 35(4), 371–388. https://doi.org/10.1023/A:1011851209975

Kelly, D., & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology,* 64(4), 745–770. https://doi.org/10.1002/asi.22799

Kemp, K. K. (2009). What can GIS offer history? *International Journal of Humanities and Arts Computing,* 3(1–2), 15–19. https://doi.org/10.3366/ijhac.2009.0006

Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., & Löfberg, L. (2016). Old Content and Modern Tools—Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910. ArXiv:1611.02839 [Cs]. http://arxiv.org/abs/1611.02839

Kramer, H. (Henk), Mücher, C. A. (Sander), & Hazeu, G. W. (Gerard). (2011). Historical land use databases: A new layer of information for geographical research. *International Journal of Humanities and Arts Computing,* 5(1), 41–58. https://doi.org/10.3366/ijhac.2011.0020

Kunz, A. (2007). Fusing Time and Space: The Historical Information System HGIS Germany. *International Journal of Humanities and Arts Computing,* 1(2), 111–122. https://doi.org/10.3366/E1753854808000219

Lifante, M. P. A., & Madrid, F. J. M. (2015). Enhancing OPAC Records: Evaluating and Fitting Within Cataloguing Standards a New Proposal of Description Parameters for Historical Astronomical Resources. *Library Resources & Technical Services,* 59(4), 140–161. https://doi.org/10.5860/lrts.59n4.140

Lin, C.-H., Hong, J.-S., & Doerr, M. (2008). Issues in an inference platform for generating deductive knowledge: A case study in cultural heritage digital libraries using the CIDOC *CRM (Vol. 8)*. https://doi.org/10.1007/s00799-008-0034-0

Mackenzie, E. S., McLaughlin, J., Moore, T. K., & Rogers, K. M. (2009). Digitising the Middle Ages: The experience of the 'Lands of the Normans' project. *International Journal of Humanities and Arts Computing,* 3(1–2), 127–142. https://doi.org/10.3366/ijhac.2009.0012

Matei, S. A. (2009). Visible Past: A location and attention aware learning and discovery environment for digital humanities. *International Journal of Humanities and Arts Computing,* 3(1–2), 163–174. https://doi.org/10.3366/ijhac.2009.0014

McEathron, S. R. (2002). Cartographic Materials as Works. *Cataloging & Classification Quarterly,* 33(3–4), 181–191. https://doi.org/10.1300/J104v33n03_09

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science,* 48(9), 810–832. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U

Othman, R., & Salahuddin, A. A. (2015). Relevance status value model of Index Islamicus on Islamic History and Civilizations. *International Journal of Web Information Systems,* 11(1), 54–86. https://doi.org/10.1108/IJWIS-06-2014-0024

Petrelli, D., & Clough, P. (2012). Analysing user's queries for cross-language image retrieval from digital library collections. *The Electronic Library,* 30(2), 197–219. https://doi.org/10.1108/02640471211221331

Porter, M. F. (2006). *An algorithm for suffix stripping. Program,* 40(3), 211–218. https://doi.org/10.1108/00330330610681286

Read, R., Taithe, B., & Ginty, R. M. (2016). Data hubris? Humanitarian information systems and the mirage of technology. *Third World Quarterly,* 37(8), 1314–1331. https://doi.org/10.1080/01436597.2015.1136208

Rodríguez, E. E., Alonso Lifante, M. P., Molero, F. J., & Randazzo, D. (2017). Advocating for a change of approach in the development of metadata standards: *Historical celestial cartography as a specialization example (Vol. 8)*. https://doi.org/10.4403/jlis.it-12398

Saiti, A., & Prokopiadou, G. (2008). Post-graduate students and learning environments: Users' perceptions regarding the choice of information sources. *International Information & Library Review,* 40(2), 94–103. https://doi.org/10.1080/10572317.2008.10762767

Saracevic, T. (2008). *Effects of Inconsistent Relevance Judgments on Information Retrieval Test Results: A Historical Perspective*. https://www.ideals.illinois.edu/handle/2142/9492

Schwartz, R. M. (2015). Digital Partnership: Combining Text Mining and GIS in a Spatial History of Sea Fishing in the United Kingdom, 1860 to 1900. *International Journal of Humanities and Arts Computing,* 9(1), 36–56. https://doi.org/10.3366/ijhac.2015.0137

Shabajee, P., McBride, B., Steer, D., & Reynolds, D. (2006). A prototype Semantic Web-based digital content exchange for schools in Singapore. *British Journal of Educational Technology,* 37(3), 461–477. https://doi.org/10.1111/j.1467-8535.2006.00616.x

Smiraglia, R. (2003). *The History of "The Work" in the Modern Catalog (Vol. 35).* https://doi.org/10.1300/J104v35n03_13

Steiner, E., & Koch, C. (2015). A Digital Archive of Cultural Heritage Objects: Standardized Metadata and Annotation Categories. *New Review of Information Networking, 20(1–2), 255–260.* https://doi.org/10.1080/13614576.2015.1112171

Suter, W. (2012). *Qualitative data, analysis, and design. Introduction to Educational Research,* 342–386. https://doi.org/10.4135/9781483384443.n12

Thiel, U., Brocks, H., Frommholz, I., Dirsch-Weigand, A., Keiper, J., Stein, A., & Neuhold, E. J. (2004). COLLATE – A collaboratory supporting research on historic European films. *International Journal on Digital Libraries,* 4(1), 8–12. https://doi.org/10.1007/s00799-003-0069-1

Ullyot, M. (2014). *Augmented Criticism, Extensible Archives, and the Progress of Renaissance Studies. Renaissance and Reformation / Renaissance et Réforme*, 37(4), 179–193.

Uzwyshyn, R. (2007). Multimedia visualization and interactive systems: Drawing board possibilities and server realities – a Cuban Rafter Paradigm Case. *Library Hi Tech,* 25(3), 379–386. https://doi.org/10.1108/07378830710820952

Webb, S., Kiefer, C., Jackson, B., Baker, J., & Eldridge, A. (2017). Mining Oral History Collections Using Music Information Retrieval Methods. *Music Reference Services Quarterly,* 20(3–4), 168–183. https://doi.org/10.1080/10588167.2017.1404307

Wettlaufer, J., Johnson, C., Scholz, M., Fichtner, M., & Thotempudi, S. G. (2015). Semantic Blumenbach: Exploration of Text–Object Relationships with Semantic Web Technology in the History of Science. *Digital Scholarship in the Humanities,* 30(suppl_1), i187–i198. https://doi.org/10.1093/llc/fqv047

Wiesman, F., van den Herik, H. J., & Hasman, A. (2004). Information Retrieval by Metabrowsing: Research Articles. *J. Am. Soc. Inf. Sci. Technol.,* 55(7), 565–578. https://doi.org/10.1002/asi.v55:7

Wilson, J. W. (2009). GIS and historical scholarship: A question of scale. *International Journal of Humanities and Arts Computing,* 3(1–2), 9–13. https://doi.org/10.3366/ijhac.2009.0005

Wohlin, C. (2014). *Guidelines for snowballing in systematic literature studies and a replication in software engineering.* EASE. https://doi.org/10.1145/2601248.2601268

Yang, C., & W. K. Luk, J. (2003). *Automatic generation of English/Chinese thesaurus based on a parallel corpus in law (Vol. 54).* https://doi.org/10.1002/asi.10259

Zeng, L., Yao, X., Liu, J., & Zhu, Q. (2014). *Construction of a one-stop document supply service platform. Interlending & Document Supply,* 42(2/3), 120–124. https://doi.org/10.1108/ILDS-01-2014-0013